LMZ Team

Project title:

Early Detection of Breast Cancer Using Gene Expression Profiling and Machine Learning Approaches

By:

1. **Lames Mohamed Ahamed Yousif Salih**
2. **Moaaz Mohammed Saadaldin Yousif**
3. **Zongo Thierry**

1. Project Overview:

The project focuses on early detection of breast cancer using RNA-Seq data, directly aligning with SDG 3: Good Health and Well-being. By leveraging machine learning, this project promotes the development of innovative diagnostic tools that enhance the quality of healthcare and improve early diagnosis, potentially saving lives.

Breast cancer is one of the leading causes of mortality worldwide due to late diagnosis and limited access to advanced diagnostic technologies. This project aims to develop a machine learning model capable of identifying biomarkers for early detection, offering a cost-effective and scalable diagnostic solution. The impact includes improved patient outcomes, reduced treatment costs, and increased awareness, contributing to the global fight against breast cancer.

2. Objectives:

• Identify Biomarkers: Use RNA-Seq data to identify gene expression patterns associated with early-stage breast cancer.
• Develop a Predictive Model: Build a machine learning model to classify breast cancer cases with high accuracy and reliability.
• Enhance Diagnostic Tools: Provide a framework that complements existing diagnostic methods, particularly in resource-constrained settings.
• Contribute to Research: Advance the field of computational biology by providing a scalable methodology for biomarker discovery.

3. Background:

Breast cancer affects millions globally, with its prognosis highly dependent on early detection. While traditional diagnostic methods rely on imaging and biopsy, these approaches are invasive, costly, and not always accessible. Machine learning offers a non-invasive alternative by analyzing RNA-Seq data to uncover biomarkers for early detection. Existing methods, such as differential gene expression analysis, have laid the groundwork, but integrating machine learning enhances the scalability and precision of biomarker discovery.

4.Methodology:

Machine Learning Techniques:

- Algorithms: Random Forest for feature selection and Support Vector Machines for classification.
- Models: Supervised learning models trained on labeled RNA-Seq datasets.
- Frameworks: Scikit-learn and TensorFlow for model development and evaluation.

Workflow Overview:

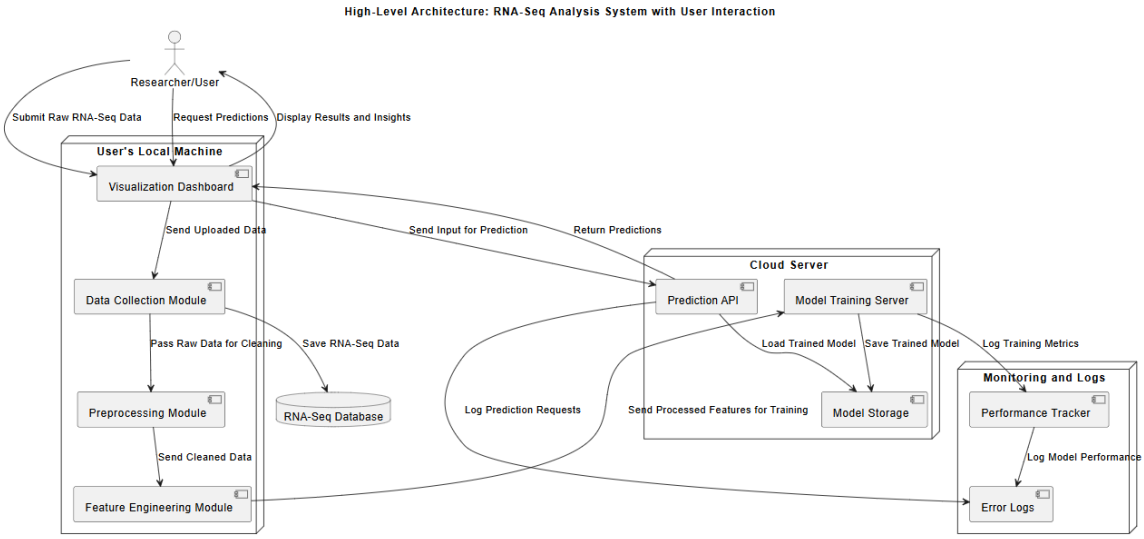
1. Data Collection: Extract RNA-Seq data from the GEO database.
2. Preprocessing: Normalize data and address class imbalance using techniques like SMOTE.
3. Feature Selection: Use Random Forest to identify key biomarkers.
4. Model Training: Train SVMs on selected features for accurate classification.
5. Validation: Evaluate models using cross-validation techniques and statistical metrics (e.g., AUC-ROC).

5. Architecture Design Diagram:

The architecture will include the following components:

1. Data Collection and Storage: Fetch RNA-Seq data from GEO, preprocess, and store in structured formats.
2. Feature Selection Module: Utilize Random Forest to identify significant genes.
3. Model Training and Testing: Train machine learning models (SVM) and evaluate their performance.
4. Visualization Dashboard: Display insights using tools like Matplotlib.

The following diagram showcases these modules in a pipeline:



6. Data Sources:

GEO Database (GSE203024): Contains RNA-Seq gene expression profiles of breast cancer patients and healthy controls. This dataset includes thousands of genes across hundreds of samples, providing a robust foundation for biomarker discovery. Preprocessing involves normalization, handling missing values, and feature scaling.

7. Literature Review:

Recent studies demonstrate the efficacy of Random Forest and SVM in cancer diagnosis. For instance, Parker et al. (2009) classified breast cancer subtypes using gene expression data. By building on these approaches, this project integrates interpretability tools like SHAP to enhance model transparency and trustworthiness.

Implementation Plan:

1. Technology Stack:

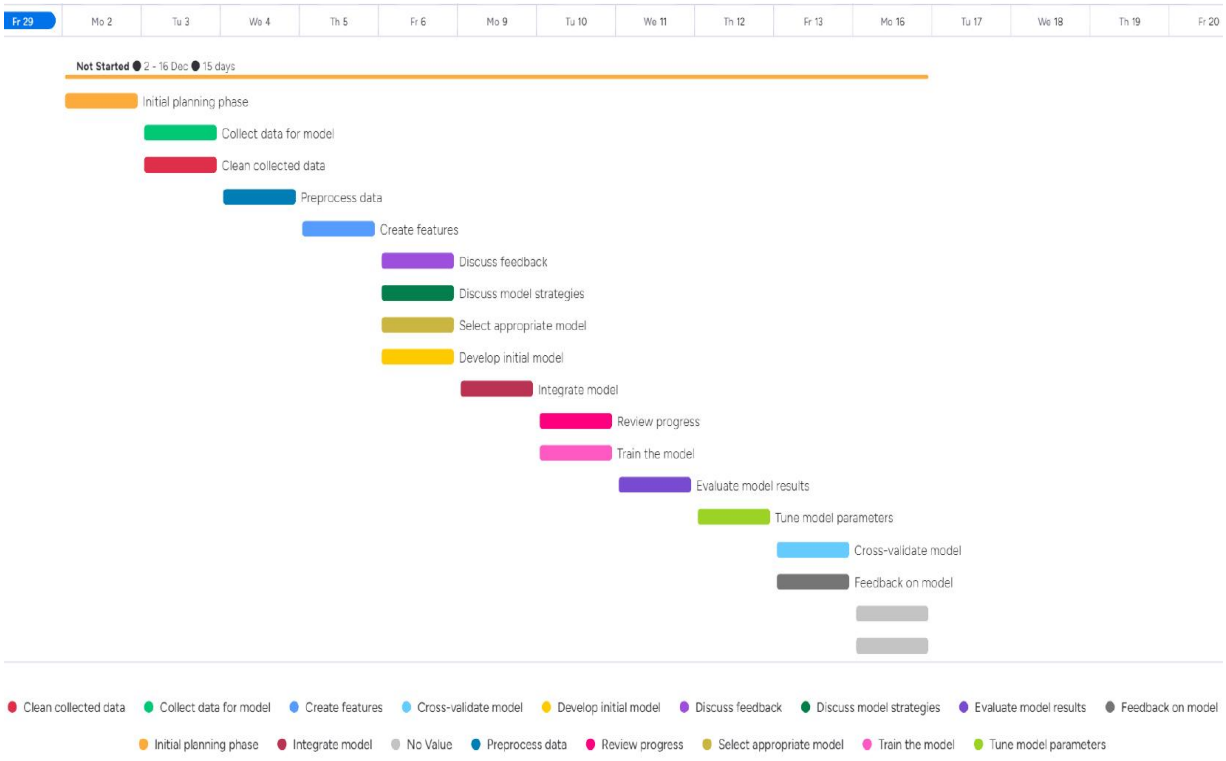
• Programming Languages: Python, R.

• Libraries/Frameworks: Scikit-learn, TensorFlow, Pandas, Matplotlib, DESeq2, Limma.

• Tools: Jupyter Notebooks.

2. Timeline:

The diagram visualizes the different stages of the project, such as data collection, model development, training and evaluation, and deployment.



Task distribution will be based on teamwork, as it's essential for learning and collective growth. We believe collaboration boosts creativity and ensures that each team member's

expertise is effectively used to achieve the project's goals. By working together, sharing ideas, and exchanging knowledge, every step becomes an opportunity for learning and mutual success.

3. Milestones:

• Completion of RNA-Seq preprocessing.
• Successful feature selection with Random Forest.
• Achieving classification accuracy >85% with SVM.
• Deployment of a functional visualization dashboard.

4. Challenges and Mitigations:

• Data Quality: Address missing values and outliers through preprocessing techniques.
• Model Performance: Employ hyperparameter tuning and cross-validation.
• Technical Constraints: Utilize cloud resources to handle large datasets.

5. Ethical Considerations:

• Data Privacy: Anonymize sensitive information from RNA-Seq datasets.
• Bias: Mitigate algorithmic bias by ensuring balanced representation of classes in training data.
• Impact: Focus on accessibility to ensure solutions benefit low-resource settings.

6. References:

1. **Random Forest for Feature Selection**
    - **Citation**: Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. DOI: 10.1023/A:1010933404324
    - **Summary**: This paper introduces the Random Forest algorithm, a popular ensemble learning method used for classification and regression tasks, particularly useful for feature selection in high-dimensional data such as gene expression profiles.
2. **Support Vector Machines (SVM) for Classification**
    - **Citation**: Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297. DOI: 10.1007/BF00994018
    - **Summary**: The foundational paper on Support Vector Machines (SVM), an effective supervised learning algorithm used for classification tasks. SVMs are well-suited for high-dimensional data like RNA-Seq due to their ability to find the optimal hyperplane for classification.
3. **DESeq2 for RNA-Seq Analysis**
    - **Citation**: Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12), 550. DOI: 10.1186/s13059-014-0550-8
    - **Summary**: DESeq2 is an R package used for differential expression analysis of RNA-Seq data. This tool uses statistical methods to normalize counts, estimate dispersion, and identify genes that show significant differences in expression between conditions.
4. **Limma for RNA-Seq Data Analysis**

- **Citation**: Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research, 43*(7), e47. DOI: 10.1093/nar/gkv007
- **Summary**: Limma (Linear Models for Microarray Data) is an R package for the analysis of gene expression data from microarrays and RNA-Seq. It provides statistical tools for differential expression analysis, making it a popular choice in bioinformatics.

5. **GEO Database for RNA-Seq Data**
   - **Citation**: Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., ... & Reinhold, W. C. (2013). NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Research, 41*(D1), D991-D995. DOI: 10.1093/nar/gks1193
   - **Summary**: The Gene Expression Omnibus (GEO) is a public functional genomics data repository that stores high-throughput gene expression data, including RNA-Seq datasets. It serves as a primary source for genomic data used in various biomedical research studies.
6. **SHAP for Model Interpretability**
   - **Citation**: Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *In Proceedings of the 31st International Conference on Neural Information Processing Systems* (Vol. 30, pp. 4765-4774). https://arxiv.org/abs/1705.07874
   - **Summary**: SHAP (SHapley Additive exPlanations) provides a unified framework for model interpretation, helping to understand how machine learning models make predictions. This is especially useful in high-stakes domains like healthcare, where model transparency is crucial.

7. **Scikit-learn for Machine Learning**

    - **Citation**: Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

    - **Summary**: Scikit-learn is one of the most widely used libraries for machine learning in Python. It provides simple and efficient tools for data mining and data analysis, including methods for classification, regression, clustering, and dimensionality reduction.